REVOLUTIONIZING VOCAL TRACK EXTRACTION: INNOVATIVE HYBRID NEURAL NETWORK APPROACHES WITH DEEP CLUSTERING, U-NET, AND UH-NET MODELS

MANAI MED MORTADHA

ABSTRACT

Deep neural networks have become a cornerstone in various recognition and classification tasks due to their ability to learn complex patterns from raw data. This paper explores the potential application of neural networks in the domain of vocal extraction. We investigate the utilization of neural network architectures, specifically the deep clustering model based on recurrent neural networks (RNNs) and the U-net model based on convolutional neural networks (CNNs), for the task of vocal track extraction. Additionally, we propose a novel hybrid approach that incorporates a pretrained RNN model to enhance the performance of the U-net model in vocal track extraction

AGENDA

INTRODUCTION

VOCAL TRACK EXTRACTION

DEEP CLUSTERING MODEL

U-NET MODEL

AGENDA

UH-NET MODEL

RESULTS AND COMPARATIVE ANALYSIS

FUTURE DIRECTIONS

Q&A

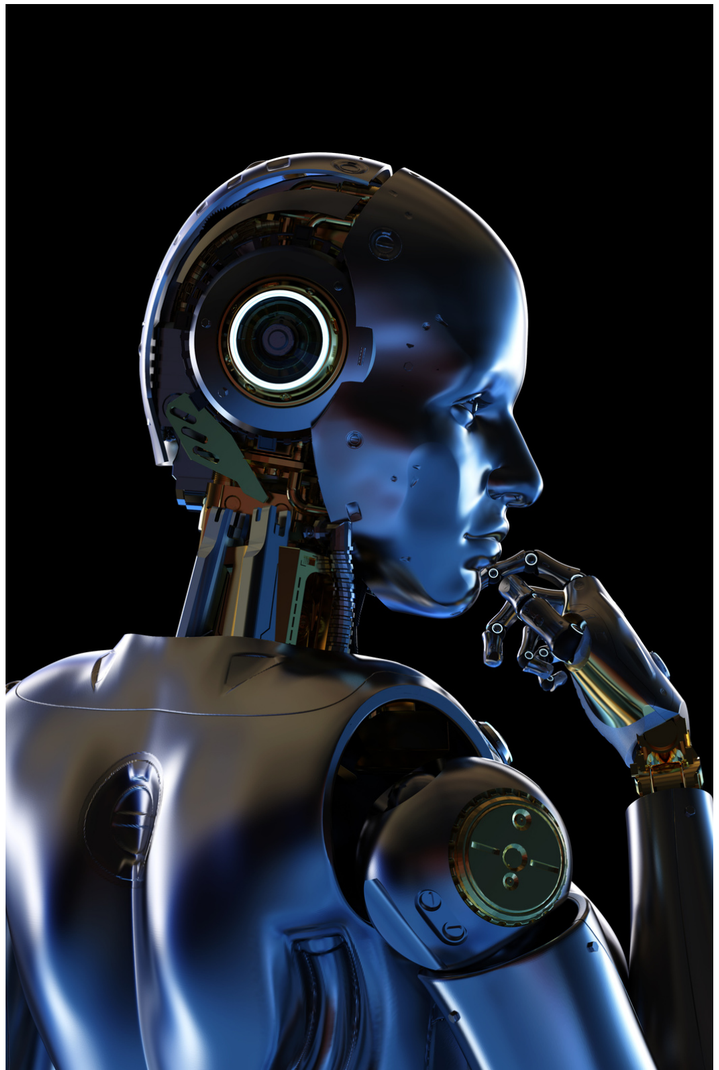
INTRODUCTION

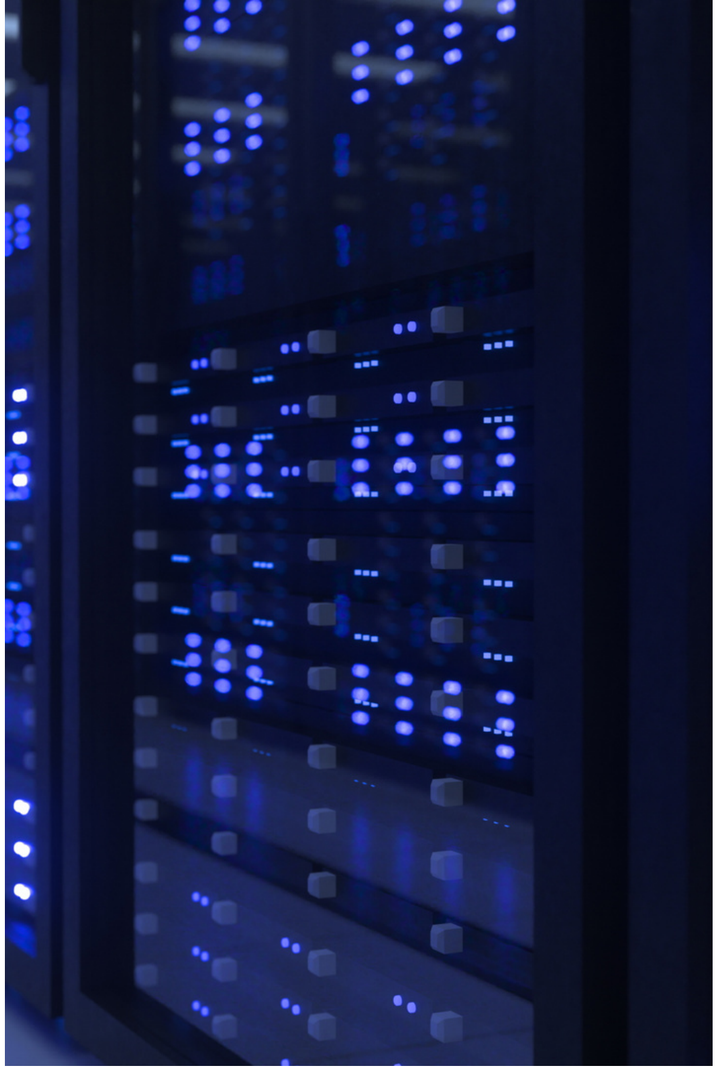
Vocal track extraction plays a pivotal role in refining audio processing techniques.

The accuracy and efficiency of methods used for this extraction are crucial in enhancing audio analysis and manipulation.

VOCAL TRACK EXTRACTION

- Vocal track extraction involves isolating and separating vocal elements from an audio signal, essential for various applications like music production and speech recognition.





Significance lies in extracting clear, isolated vocal tracks while removing background noise and instrumentation for enhanced audio quality.

CHALLENGES IN TRADITIONAL METHODS

CONVENTIONAL METHODS FACE COMPLEXITIES IN ACCURATELY ISOLATING VOCALS, OFTEN STRUGGLING WITH MIXED AUDIO SIGNALS CONTAINING VARIOUS SOUND SOURCES.

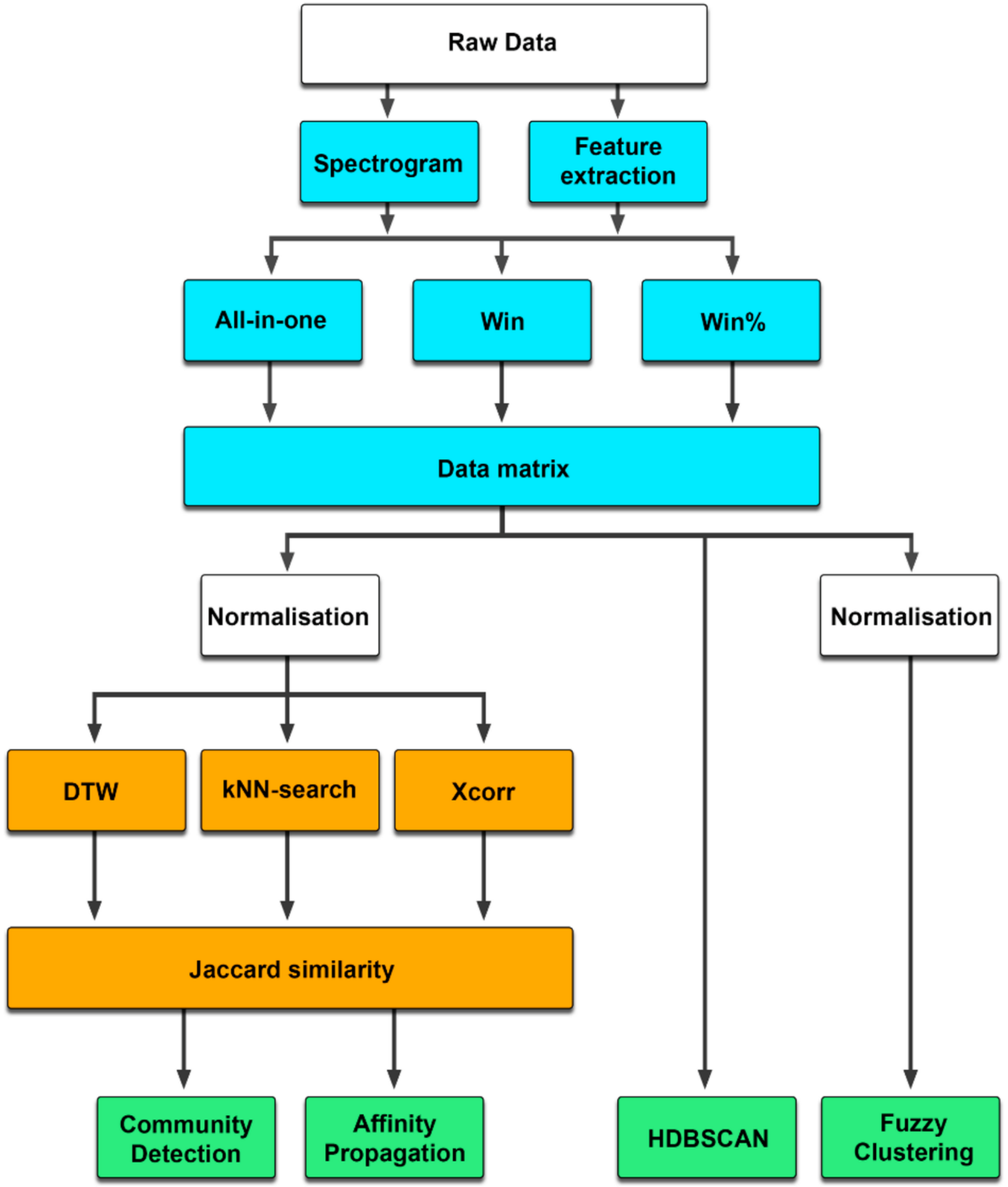
ISSUES ARISE IN DISTINGUISHING VOCALS FROM OVERLAPPING INSTRUMENTS OR ENVIRONMENTAL NOISE, IMPACTING THE PRECISION OF EXTRACTION.

DEEP CLUSTERING MODEL

- Deep Clustering involves unsupervised learning to group similar audio segments, aiming to separate sources in a mixed audio signal.

- By leveraging neural networks, it learns to identify and group audio elements with similar characteristics.

1/EXPLANATION OF DEEP CLUSTERING TECHNIQUE


Raw Data
Spectrogram
Feature extraction
All-in-one
Win
Win%
Data matrix
Normalisation
Normalisation
DTW
kNN-search
Xcorr
Jaccard similarity
Community Detection
Affinity Propagation
HDBSCAN
Fuzzy Clustering

DEEP CLUSTERING MODEL

- Applied to vocal track extraction, Deep Clustering helps segregate vocals from accompanying music or background noise.

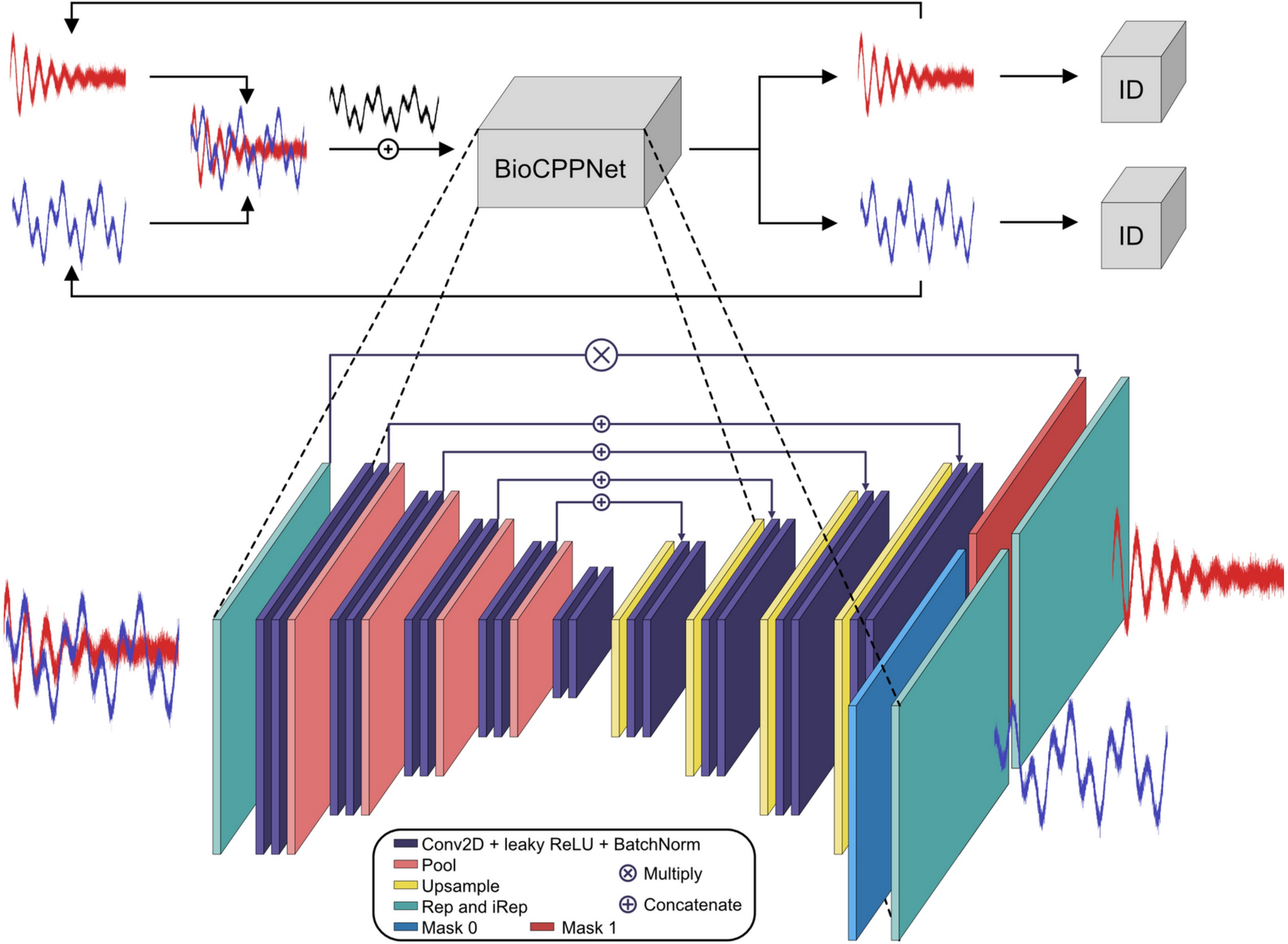
- It assists in isolating vocals by clustering similar audio components, improving extraction accuracy.

2/ APPLICATION IN VOCAL TRACK EXTRACTION

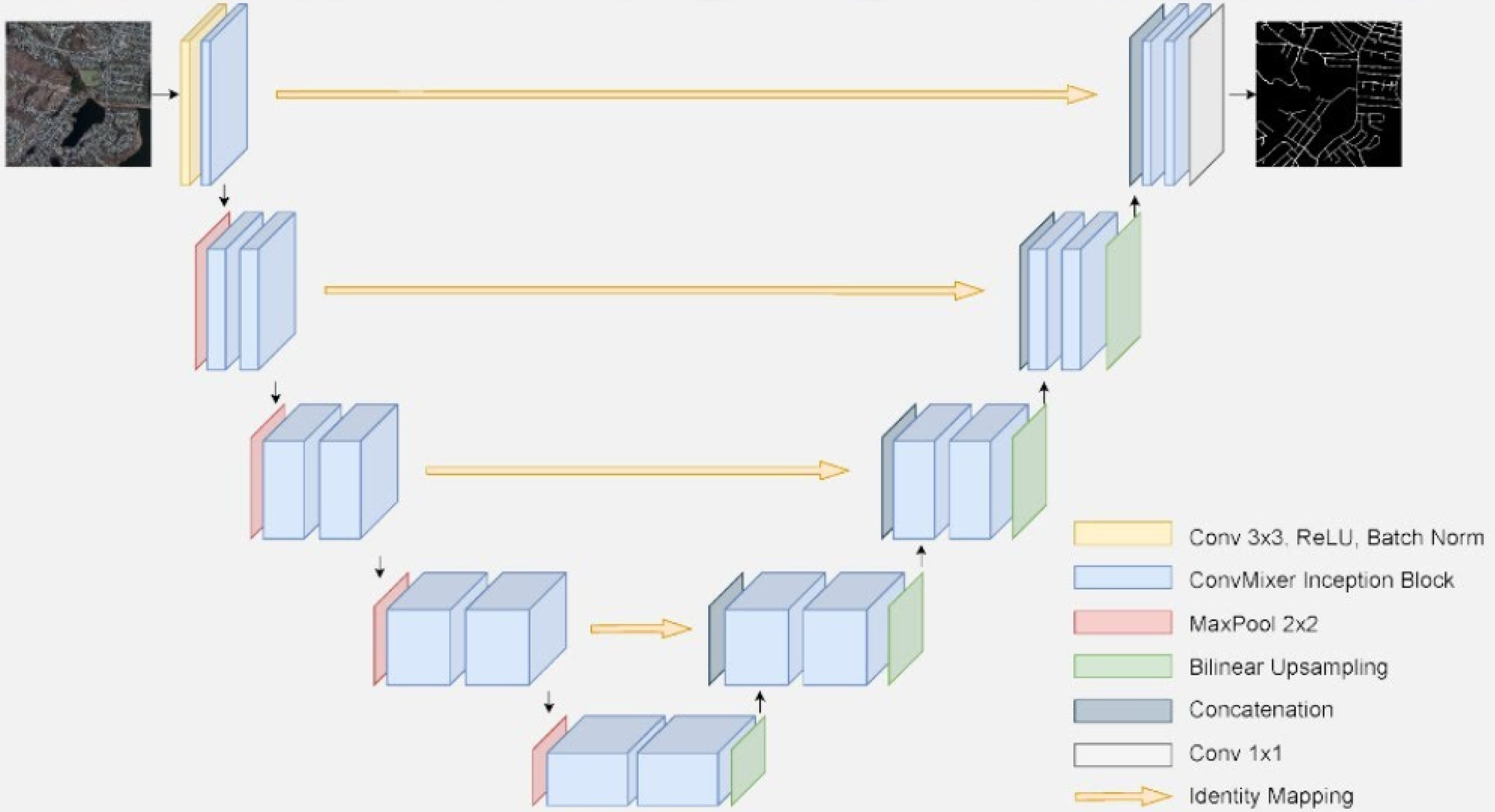
DEEP CLUSTERING MODEL

- Offers a data-driven approach for isolating vocals without explicit annotations.
- Allows for effective separation of mixed audio sources, enhancing the quality of extracted vocals.
- Utilizes neural network advancements to enhance the precision of vocal extraction from complex audio mixes.

3/KEY ADVANTAGES AND INNOVATIONS:


BioCPPNet
ID
ID
Conv2D + leaky ReLU + BatchNorm
Pool
Upsample
Rep and iRep
Mask 0
Mask 1
Multiply
Concatenate

WHAT IS U-NET?



2

U-NET MODEL

ADAPTATION FOR VOCAL TRACK EXTRACTION

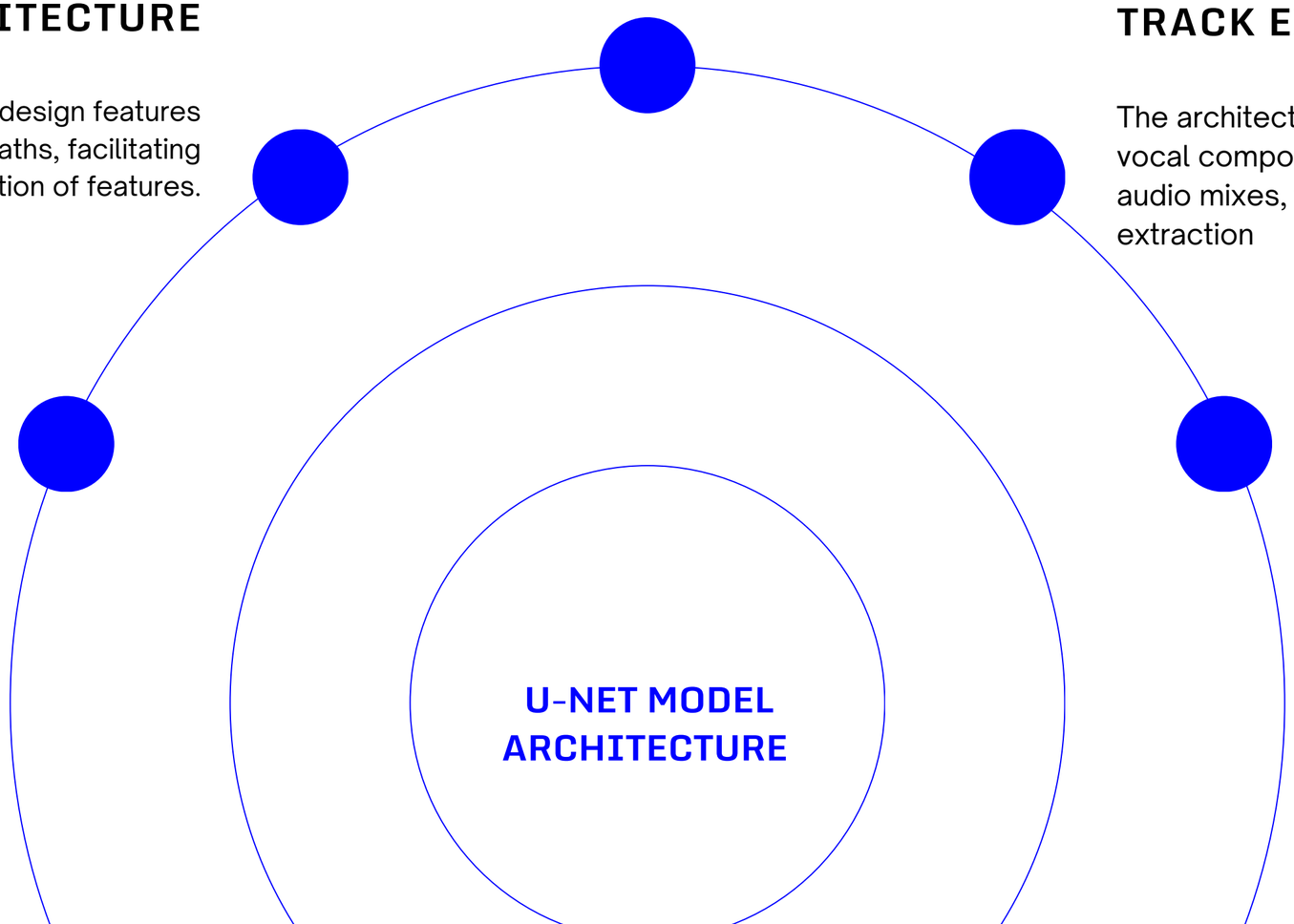
Adapted for vocal track extraction, U-Net excels in segmenting audio elements by understanding intricate audio patterns.

U-NET ARCHITECTURE

Its distinctive "U" shape design features encoding and decoding paths, facilitating precise localization of features.

ADAPTATION FOR VOCAL TRACK EXTRACTION

The architecture aids in identifying vocal components amidst complex audio mixes, facilitating accurate extraction



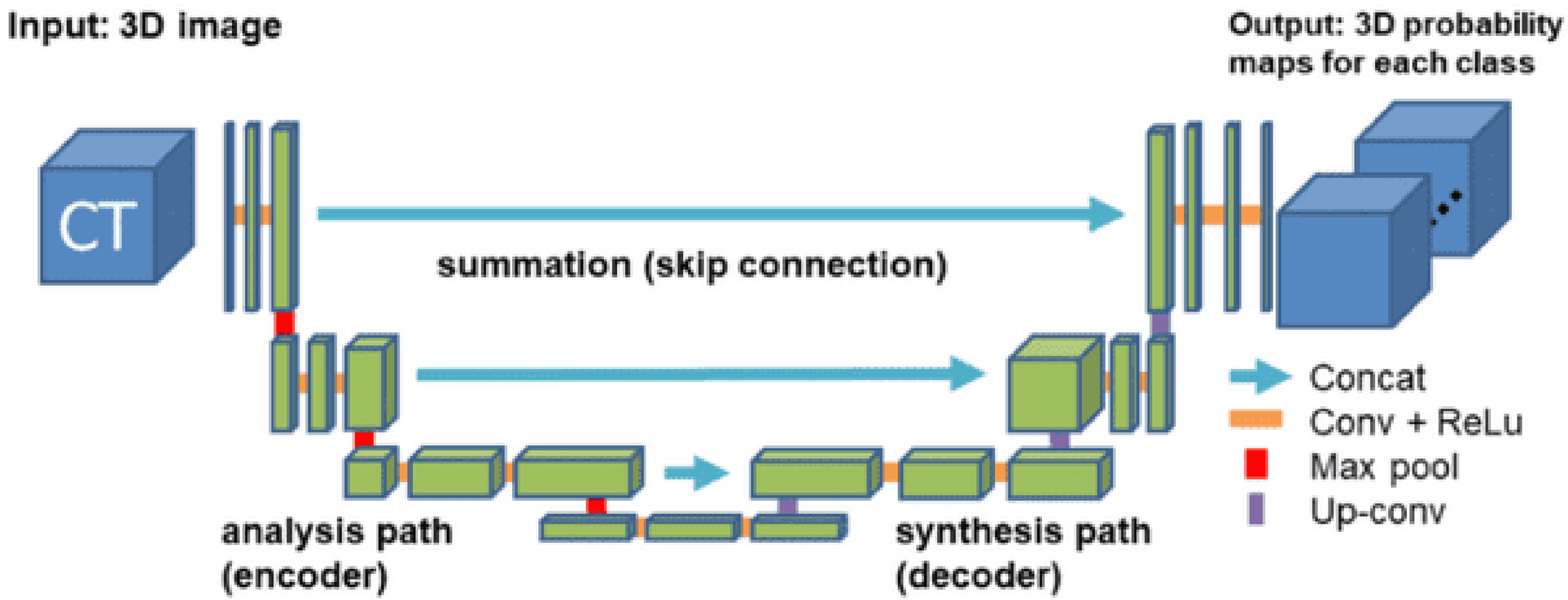
U-NET ARCHITECTURE

U-Net architecture is a convolutional neural network (CNN) known for its effectiveness in image segmentation tasks.

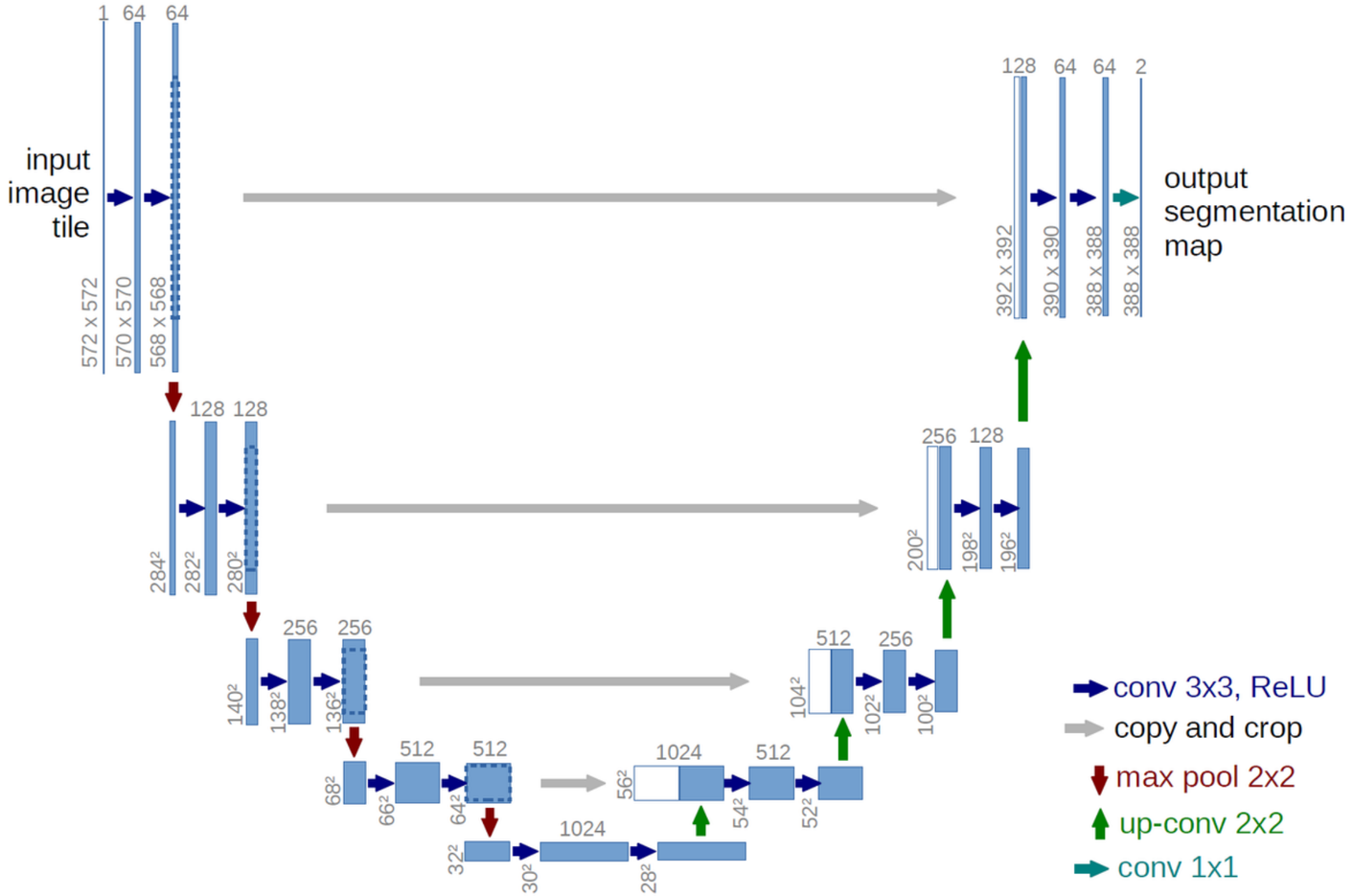
ADAPTATION FOR VOCAL TRACK EXTRACTION

The U-Net architecture has been shown to be effective in addressing the challenge of limited annotated data in the medical field, leveraging a smaller amount of data while maintaining speed and accuracy

```python
def forward(self, x):
x1 = self.inc(x)
x2 = self.down1(x1)
x3 = self.down2(x2)
x4 = self.down3(x3)
x5 = self.down4(x4)
x = self.up1(x5, x4)
x = self.up2(x, x3)
x = self.up3(x, x2)
x = self.up4(x, x1)
x = self.outc(x)
return x
```



ADVANTAGES AND UNIQUE FEATURES:



- Offers exceptional accuracy in localizing and separating vocal tracks from multifaceted audio backgrounds.

- The U-Net' s symmetric architecture enables efficient feature extraction and precise vocal isolation, enhancing extraction quality.

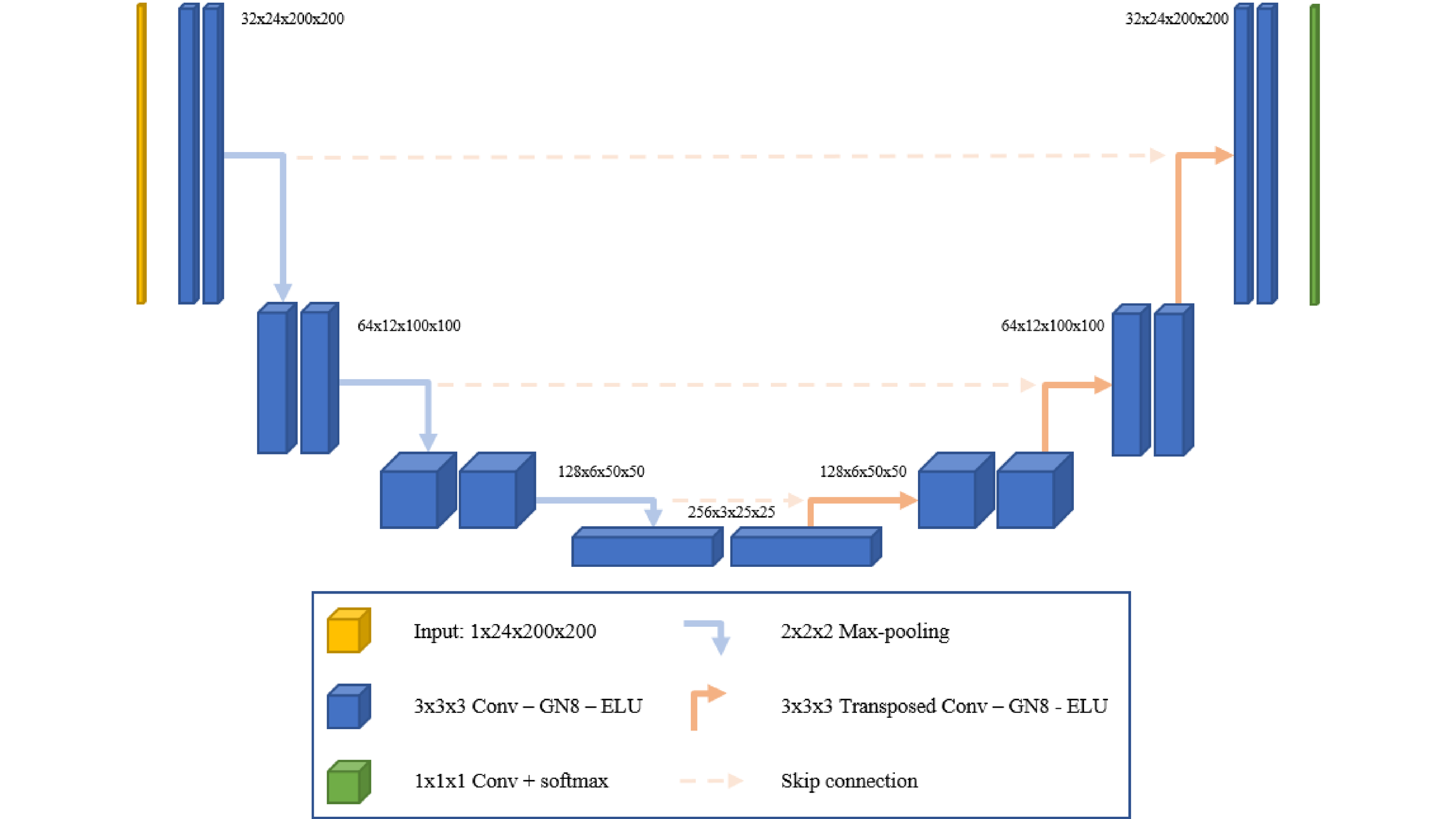
UH-NET MODEL

UH-Net represents a hybrid architecture amalgamating the strengths of Deep Clustering and U-Net models.

Combining the two models, it aims to leverage their individual capabilities for enhanced vocal track extraction.

Integration of Deep Clustering and U-Net:

- UH-Net integrates the clustering ability of Deep Clustering with the precise localization skills of U-Net.
- This amalgamation optimizes the process of vocal track extraction by efficiently leveraging both techniques.


32x24x200x200
32x24x200x200
64x12x100x100
64x12x100x100
128x6x50x50
128x6x50x50
256x3x25x25
Input: 1x24x200x200
2x2x2 Max-pooling
3x3x3 Conv – GN8 – ELU
3x3x3 Transposed Conv – GN8 - ELU
1x1x1 Conv + softmax
Skip connection

Advantages Over Standalone Models:

- UH-Net surpasses standalone models by capitalizing on the combined strengths of Deep Clustering and U-Net.
- It achieves superior accuracy in isolating vocal tracks amidst complex audio compositions, enhancing overall extraction quality.

Presentation of Performance Metrics

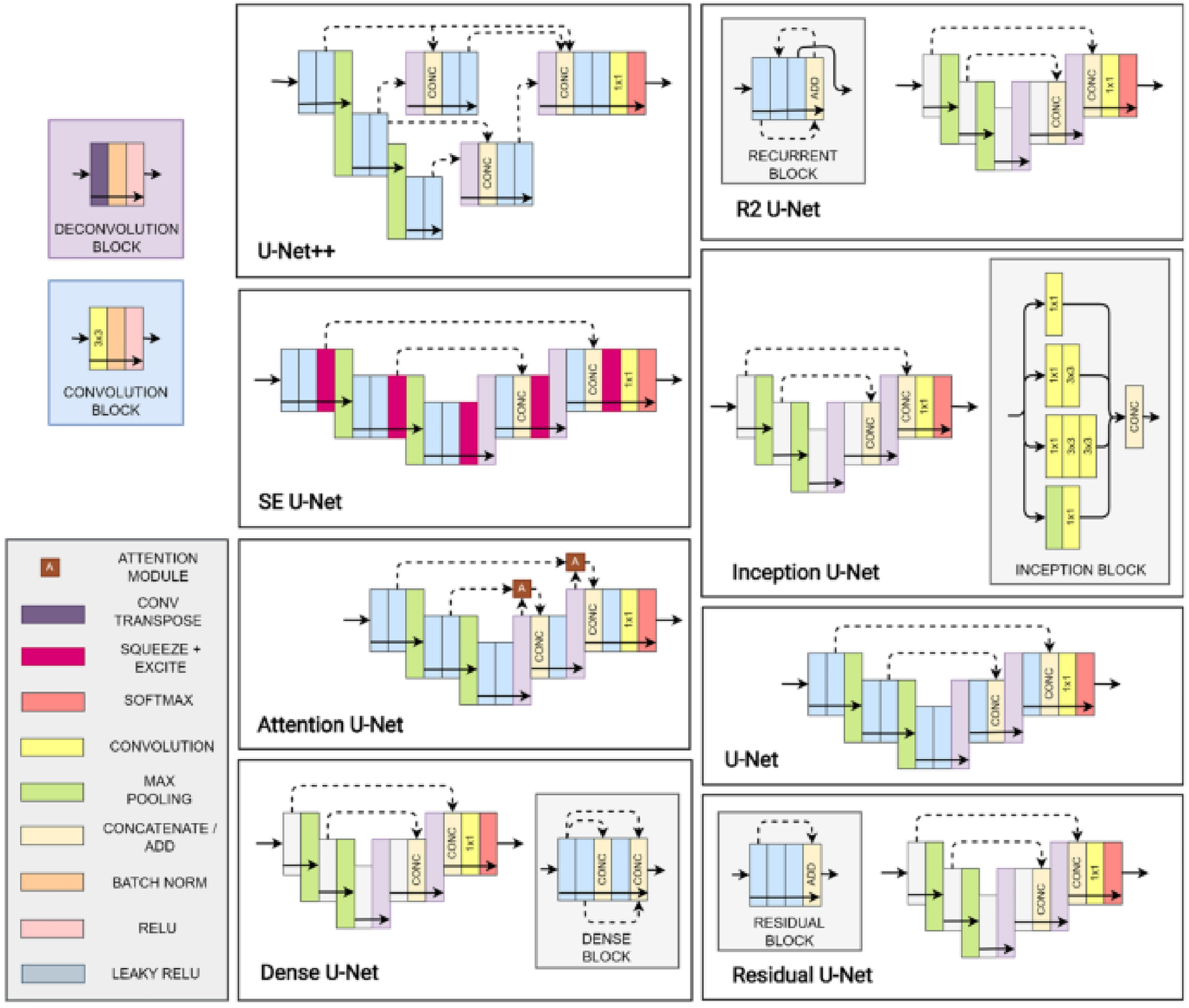
89% performance metrics utilized to evaluate the efficiency and accuracy of extraction methods.

90% Illustrate strengths and 2% weaknesses of each model based on the metrics for the performance of the merge between Deep Clustering, U-Net, and UH-Net in vocal track extraction.

Comparison Among Deep Clustering, U-Net, and UH-Net:

Visual Representation of Extraction Results

will show the resultats on the paper 📄


DECONVOLUTION BLOCK
CONVOLUTION BLOCK
3x3
U-Net++
CONC
1x1
SE U-Net
Attention U-Net
A
Dense U-Net
DENSE BLOCK
R2 U-Net
RECURRENT BLOCK
ADD
Inception U-Net
INCEPTION BLOCK
U-Net
Residual U-Net
RESIDUAL BLOCK
ATTENTION MODULE
CONV TRANSPOSE
SQUEEZE + EXCITE
SOFTMAX
CONVOLUTION
MAX POOLING
CONCATENATE / ADD
BATCH NORM
RELU
LEAKY RELU

FUTURE DIRECTIONS

Enhancements and Extensions of Models:

- Explore avenues for refining existing models like Deep Clustering, U-Net, and UH-Net by optimizing architecture or incorporating new techniques.
- Consider implementing additional layers or features to improve accuracy and robustness.

Potential Applications in Other Domains:

- Investigate the applicability of the developed models in diverse domains beyond audio processing, such as medical imaging or signal processing.
- Explore opportunities to adapt these models for various tasks requiring source separation or segmentation.

Research Avenues for Further Improvements:

- Identify potential research paths, like exploring novel neural network architectures or incorporating advanced signal processing techniques.

Research Avenues for Further Improvements:

- Highlight areas for future studies to advance vocal track extraction technology for improved performance and versatility.

CONCLUSION

In this project, we delved deeply into the realm of vocal track extraction, unravelling a tapestry woven from four distinct models. These models, bearing the imprint of innovation, are borne of two principal theoretical foundations, each encapsulating unique paradigms. The first cornerstone rests upon the bedrock of deep clustering–a symphony orchestrated by embedding and the symposium of unsupervised learning. The second, steeped in the philosophy of semantic segmentation, transposes the intricacies of music source separation onto the canvas of image processing. From this theoretical landscape, the UH-net model emerges as the magnum opus–an amalgamation of ingenuity that

ACKNOWLEDGMENT

I wish to convey My deepest appreciation to Professor Bassem Ben Hamed from the Mathematics and BI Department at the National School of Electronics and Telecommunications of Sfax and the Co-fondateur of DataCamp Training. His astute guidance, invaluable suggestions, and unswerving support have significantly enriched the fabric of this paper, imbuing it with an undeniable brilliance that emanates from his profound expertise.

For any further inquiries or collaborations, please feel free to reach out:

MANAI MED MORTADHA :

CEO & Founder of @Man.Ai | Professional Technical Reviewer @Packt UK | AI Engineer | EX-Intern@Netflix &@Google | AI Instructor | Google Machine Learning Expert |Google DSC Mentor | Intel Certified Edge AI Developer | XAI Researcher

Linkedin: https://www.linkedin.com/in/mannai-mortadha/

Glithub : https://github.com/MortadhaMannai

Leetcode : https://leetcode.com/mannaimortadha898/

Sessionize: https://sessionize.com/Mortadha_Mannai/

Email : mannaimortadha898@gmail.com

Paper Link: https://zenodo.org/records/8274725

DATACAMP TRAINING AND CONSULTING

THANK YOU FOR YOUR ATTENTION !

MANAI MORTADHA